DOCUMENT RESUME

ED 423 290 TM 029 097

AUTHOR Sykes, Robert C.; Ito, Kyoko
TITLE The Effect of Rounding Aggregated Item Ratings for Constructed Response Items in Mixed-Item Format Tests.
PUB DATE 1998-04-00
NOTE 35p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (San Diego, CA, April 14-16, 1998).
PUB TYPE Reports - Research (143) -- Speeches/Meeting Papers (150)
EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Ability; *Constructed Response; Judges; *Scoring; *Test Format; Test Items
IDENTIFIERS *Aggregation (Data); High Stakes Tests; *Rounding (Mathematics)

ABSTRACT

A common procedure for obtaining multiple readings (ratings) for a constructed response item, especially in high-stakes tests, is to have two readers read the papers independently, with a third reading if the results differ by more than one point. This necessitates a scoring rule that specifies how the ratings will be aggregated into a single item score. Two plausible scoring rules involve averaging the readings and rounding either to the nearest half point or the nearest integer, but it is not known which results in a greater precision of measurement. This study investigated the precision and accuracy of ability estimates obtained under the two scoring rules for mixed format tests calibrated under an item response theory model. Eleventh-grade reading, mathematics, and science test results and a fifth-grade mathematics test result were analyzed, with more than 1,200 students available for each form. There was little substantive difference in score information or the standard errors of ability estimates due to the type of rounding (integer versus half point), above the floors of three of the four tests, but in the fourth (11th grade reading) there was less error in the integer-rounded ability estimates at the lower portion of the scale. Integer-rounded estimates generally produce slightly larger predicted percent of maximum (test) scores, though not throughout the entire ability range of all the four tests studied. The expected larger positive differences or rounding bias for number correct estimates were observed. Within-subject differences between scale score estimates derived using integer versus half-point scores were generally small for both pattern and number correct ability estimates. The lack of substantive improvement in measurement precision that could be attributed to half-point rounding, coupled with the documented instance of increased error induced by that type of rounding in a portion of the ability range of students taking one test, would seem to argue for rounding average ratings to the nearest integer. Rounding up gives the preponderance of students the benefit of the doubt concerning the acceptability of their responses. (Contains two tables, four figures, and eight references.) (SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED 423 290

The Effect of Rounding Aggregated Item Ratings for Constructed Response Items in Mixed-Item Format Tests

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY

Robert Sykes

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

1

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

Robert C. Sykes

Kyoko Ito

CTB/McGraw-Hill

This paper was presented at the Annual Meeting of the National Council on Measurement in Education in San Diego, April 1998.

TM029097

BEST COPY AVAILABLE

INTRODUCTION

A common procedure for obtaining multiple readings (ratings) for a constructed response (c.r.) item, particularly those c.r. items in tests used to make high-stakes decisions, is to have two readers independently read the papers with a third independent reading acquired if the ratings differ by more than one point. The presence of two or three readings of a response to a c.r. item necessitates a scoring rule that specifies how the ratings (readings) will be aggregated into a single item score.

Two plausible scoring rules involve averaging the two (or three) item ratings and rounding either to the nearest half point or to the nearest integer. Both rules are compatible with tests containing multiple item types (mixed-format tests incorporating multiple choice {m.c.}and c.r. items) that are scaled using a generalized IRT model incorporating a three-parameter logistic model (3pl) for the m.c. items and a two-parameter partial credit model (2ppc) for the c.r. items. This 3pl/2ppc type of generalized IRT model has been shown to better fit items in mixed-format tests (Fitzpatrick, Link, Yen, Burket, Ito & Sykes, 1996).

It is not known whether increasing the number of levels by rounding to the nearest half point results in greater precision of measurement than rounding to the nearest integer. Half point item scores would reflect rater disagreement. A potential drawback to rounding to the nearest integer is that for the majority of a student's responses to the c.r. items only two readings will be necessary. A one point disagreement would always result in an

average item score that is rounded up, introducing varying degrees of positive "rounding" bias into the total raw scores.

The precision or reliability of "half-point round" versus "integer-round" c.r. item scores on ability estimates can be assessed by the evaluation of information functions for the composite test scores to which they contribute. The reciprocal of the information function for a composite score at a particular ability level is the standard error of ability estimate (s.e.). If the use of half score points increases measurement precision (reduces error), test scores utilizing them should demonstrate lower s.e.s across at least portions of the ability range.

Differences in information may result from differences in how ability estimates weight component item scores. The weighting by item discrimination associated with pattern scoring, which utilizes the examinee's pattern of responses to the items, allows an item's contribution to an ability estimate to vary relative to the degree to which item scores are associated with ability. Conversely number-correct scoring, by considering one item point or level to be as good as any other, requires that each point contribute equally to the total score and derived ability estimate.

The degree of rounding *bias* that is incurred by rounding the average readings for a c.r. item to the nearest integer may be evaluated by comparing the test characteristic curves (tccs) for ability estimates derived using half-point-rounded scores with tccs obtained using integer-rounded scores. When the tccs are

obtained by scoring a single sample both ways, differences in the expected number correct or predicted percentage of maximum score (predicted pm) for integer-rounded c.r scores relative to the prediction of the tcc for half-point-rounded scores reflects the bias or inaccuracy due to type of rounding.

The purpose of this research was to investigate the precision and accuracy of ability estimates obtained under the two scoring rules for mixed-format tests calibrated with a 3pl/2ppc IRT generalized model. Test information and tccs obtained through the application of the two scoring rules were compared for each of two types of ability estimates: pattern and number-correct scores. Additionally within-subject differences in examinees' scaled scores were evaluated for signs that subsets of examinees may be substantively advantaged or disadvantaged by the manner of rounding employed.

METHOD

Source Data

Mixed-format pilot (operational forms undergoing a final pre-operational administration) eleventh grade Reading, Math, and Science forms and a single tryout fifth grade Math form were available from two testing programs. The number of scored items of each type and the range in the number of levels (including 0) of the c.r. items are summarized below:

Content Area	Grade	Multiple Choice	Constructed Response	Range in Number of Levels of C.R. Items: Half-Point-Rounded	Range in Number of Levels of C.R. Items: Integer-Rounded
Reading	11	35	1	(7 - 7)	(4 - 4)
Math	11	40	6	(5 - 11)	(3 - 6)
Science	11	42	8	(5 - 7)	(3 - 4)
Math	5	49*	11	(5 - 9)	(3 - 5)

Eighteen of the 49 items denoted as multiple choice for the Math/Grade 5 test were actually gridded response items. Although similar to the 11 three to five level c.r items in their being scaled with a partial credit model, they are not considered c.r. items for the purpose of this study because their scores do not involve ratings or their averaging. More than 1200 students were available for each form.

Rating Process

Each c.r. item in the four tests was scored by at least two readers. If the readers' scores differed by more than one point, a third rating was obtained. Half-point-rounded scores were obtained by averaging the two or three ratings for an item and rounding to the nearest half point. Integer-rounded c.r. item scores resulted from rounding the average rating to the nearest integer.

The implemented rating process resulted in the production of four meaningful kinds of averages. An average score equal to an integer could occur with either two or three readings. A second kind of average consisted of a score with a remainder of ½ when two readers disagreed by a single point. The final two kinds or types of averages occurred when the average of three readings had

BEST COPY AVAILABLE

a remainder of 1/3 or 2/3.

Averages with a remainder of ½ or 2/3 would be rounded up to the next integer with integer-rounding (e.g. 2.5 or 2.67 to 3.0). An average with a remainder of 1/3 would be reduced to the lower integer with this type of rounding. Average scores with any of the three possible remainders would be rounded to the half point with half-point rounding.

Readers for both testing programs were trained to implement scoring rubrics; anchor papers, check sets, and read behinds were employed to verify and maintain scoring accuracy. Inter-rater reliability studies that incorporated second reads for a large sample of students taking each test indicated that the percentage of exact agreement on the 15 c.r. items in the three eleventh-grade tests ranged from 68% to 93%. Minimum and maximum exact agreement rates of 51% and 97% were obtained in a similar manner for the 11 c.r. items in the fifth grade Math test. Approximate agreement (within one point) ranged between 89% and 100% across the c.r. items in all four tests.

Scaling Process

Multiple-choice and open-ended items were scaled together twice using the generalized IRT model. With the generalized model a three-parameter logistic model (Lord, 1980) was used for the multiple-choice items:

$$P_i = P(X_i = 1|\theta) = c_i + \frac{1 - c_i}{1 + \exp[-1.7A_i(\theta - B_i)]}, \qquad (1)$$

where A_i is the discrimination, B_i is the difficulty, and c_i is the lower asymptote or guessing parameter for item i.

A generalization of Master's (1982) Partial Credit model was used for the c.r. items. This 2PPC model is the same as Muraki's (1992) "generalized partial credit model." For a c.r. item with m_i score levels assigned integer scores that ranged from 0 to m_i - 1:

$$P_{ik}(\theta) = P(X_i = k-1|\theta) = \frac{\exp(y_{ik})}{\sum_{j=1}^{m_i}\exp(y_{ij})}, \qquad k = 1,\ldots,m_i \tag{2}$$

where

$$y_{ik} = \alpha_i(k-1)\theta - \sum_{j=0}^{k-1}\gamma_{ij},$$

and $\gamma_{i0} = 0$. α_i is the item discrimination. γ_{ij} is related to the difficulty of the item levels: the trace lines for adjacent score levels intersect at $-\gamma_{ij}/\alpha_i$.

Parameter Estimation and Model Predictions of Performance

Item parameter and θ estimation was conducted using the program PARDUX (Burket, 1991; 1995). Item parameters were estimated using marginal maximum likelihood procedures implemented with an EM algorithm. Evaluations of the accuracy of the program with simulated data (Fitzpatrick, 1994) have found it to be at least as accurate as MULTILOG (Thissen, 1986). The ability scale was defined by specifying a prior true θ distribution to have a mean of 0.0 and standard deviation of 1.0.

Maximum likelihood ability estimates were obtained. For reporting purposes, the ability estimates obtained for each test were linearly transformed to a scale score metric by multiplying by 50 and adding 500. The pattern, though not the number correct scale scores, that resulted were expressed to the half point.

Fit

Model fit was evaluated with a generalization of the Yen (1981) Q_1 statistic comparing observed and predicted trace lines (Fitzpatrick, et al., 1996). The Fit z is a standardization of the Q_1 statistic that facilitates comparisons of items with varying numbers of score levels :

$$Z_{Q1} = \frac{Q_1 - df}{\sqrt{2df}}. \qquad (3)$$

The power of z increases with sample size, so for flagging purposes the statistic is typically compared to critical values that increase with the size of samples. For samples of the size used in this study a value of 4.0 was used to flag items for misfit.

Observed and predicted trace lines were also compared graphically. Because of the difficulty of interpreting multiple trace lines plots for multi-level items, observed item performance was compared against predicted performance using the item characteristic function:

$$E(X_i|\theta) = \sum_{k=1}^{m} (k-1) P_{ik}(\theta). \qquad (4)$$

Predictions of test performance were made through test characteristic functions obtained by summing item characteristic functions:

$$E(X_t / n_{t\,mxsc} | \theta) = [\sum_{i=1}^{nitem} \sum_{k=1}^{m_i} (k-1) P_{ik}(\theta)] / n_{t\,mxsc} ,$$

where X_t is the test score and $n_{t\,mxsc}$ is the maximum number of points in the test. After multiplication by 100 a predicted percentage of maximum test score was obtained.

Information

Results were evaluated with respect to test score information. The pattern scores produced using the 3PL/2PPC model utilizes optimal scoring weights, w_i, which maximize test information. For the 3PL items, these weights are defined by Lord (1980, Section 4.13), and for the 2PPC items the optimal weights are α_i. When the optimal weights are used, the test score information is the sum of item information functions, defined as:

$$I(\theta, X_t) = \sum_{k=1}^{m_i} \frac{[P'_{ik}(\theta)]^2}{P_{ik}(\theta)} , \qquad (5)$$

where $P'_{ik}(\theta)$ is the derivative of $P_{ik}(\theta)$ with respect to θ. Test score information is subsequently:

$$I(\theta, \sum_i w_i X_i) = \sum_{i=1}^{n} \sum_{k=1}^{m_i} \frac{[P'_{ik}(\theta)]^2}{P_{ik}(\theta)} . \qquad (6)$$

It is also possible to base the ability estimate on an unweighted sum of item scores.[In fact it is possible to base the

trait estimate on any arbitrary set of item weights.] Following the logic of Lord (1980, Equation 5-3), the information of the unweighted raw score is

$$I(\theta, \sum_i X_i) = \frac{\left[\sum_{i=1}^{n}\sum_{k=1}^{m_i}(k-1)P'_{ik}(\theta)\right]^2}{\sum_{i=1}^{n}\sigma^2(X_i|\theta)}. \qquad (7)$$

For a given model, the information in the unweighted raw score is less than or equal to the information of the optimally weighted score.

Evaluation of Score Precision and Bias Due to Rounding

The error associated with integer-rounded scores was evaluated through comparisons of model predictions, specifically test score information/standard errors of ability estimates and predicted pm's, as well as comparisons of ability estimates obtained from a single sample of students taking each of the four tests. Consequently student responses to the items of a test differed across the type of rounding condition only in the manner in which ratings for the c.r. items were rounded (responses to all other items were identical). The two types of rounded c.r. item scores were each utilized in the estimation of a pattern and number-correct test score, resulting in four combinations of type of c.r. item score (half-point versus integer) and ability estimate (pattern score versus number-correct).

RESULTS

Raw Score Statistics

Descriptive statistics for the four tests are presented in Table 1. The three grade 11 tests were moderately difficult, with means expressed as observed percent of maximum scores ranging between 50% and 57%. The Math/Grade 5 test was more difficult, with students, on average, obtaining 35% of the total 75 points when either integer-rounded or half-point-rounded c.r. scores were used.

The mean of the total scores containing the integer-rounded c.r. item scores was, as expected, higher than the mean total score containing half-point c.r. item scores for each of the four tests. The increase ranged between .13 (21.73 - 21.60) for the Reading/Grade 11 test with its single c.r. item to .79 for the Science/Grade 11 test with its eight c.r. items. The difference in total scores is attributed solely to the differences in c.r. scores induced by the type of rounding (e.g. the difference of .13 between the mean integer-rounded c.r. score and the mean half-point-rounded c.r. score {.79 versus .66, respectively}).

Scaling Results

All items in each of the four forms were calibrated twice with the 3pl/2ppc model, once using half-point-rounded c.r. item scores (and student responses to all other items) and the other time with integer-rounded scores.

The largest number of misfitting items within the eight item calibrations (two types of rounding times four tests) was six for

the half-point-rounded Math/Grade 5 calibration (average z=6.53), followed by five for the integer-rounded estimates for the same tryout form (average z = 6.37). The largest absolute difference between a predicted and observed p-value for these 11 misfitting sets of item parameter estimates was .006.

Two of the remaining six item calibrations (integer-rounded item parameter estimates for Math/Grade 11 and half-point-rounded estimates for Reading/Grade 11) had two misfitting items each and the other four had only a single misfitting item. The largest absolute deviation between observed and predicted p-values for the misfitting items in these six calibrations was .01.

No c.r. item, when calibrated with integer-rounded or half-point-rounded c.r. scores, misfit.

Information

Figures 1 through 4 contain plots of the score information functions of the four combinations of ability estimate by type of rounded c.r. item score for Reading/Grade 11, Science/Grade 11, Math/Grade 11, and Math/Grade 5, respectively. Presented below the plots of information are the reciprocal values of the four score information functions, the standard error of ability estimates for scale score intervals of 25 points between 300 and 700, inclusive. A frequency distribution of the number correct, or unweighted half-point-round ability estimates, permits an assessment of relatively how many examinees falls at each of the scale score values.

Pattern Ability Estimates

All four score information plots reveal that pattern estimates provide the most information(and least error), regardless of type of rounding. This is expected, due to the greater efficiency of pattern scoring. The plots of score information for the pattern scores are very nearly coincident for all four tests. An evaluation of the tabled s.e.'s indicates that, with the exception of the lower portion of the scale score range for Reading/Grade 11 (up through 425), the difference between the s.e.'s for integer vs half-point pattern ability estimates is no more than five points, and very frequently no more than two points. Hence, there appears to be no substantive difference in the precision of the two types of scores through most of the score ranges for the four tests.

The lower portion of the scale score range for Reading/Grade 11 demonstrates markedly smaller s.e.'s for the integer-rounded pattern scale scores relative to the half-point-rounded scores, however. At the floor for this test, a scale score of 300, the integer-rounded s.e. is 69 points less than that for the half-point s.e. (132 versus 201) and remains 14 points less at a scale score of 400. The 69 point difference at the floor is larger than one integer-rounded or half-point-rounded pattern score standard deviation (approximately 64 scale score points - Table 2). The greater precision of the integer-rounded pattern ability estimates implies that half point scores actually degrade the precision of measurement in this subrange.

Number Correct Ability Estimates

An evaluation of the information provided by integer-rounded versus half-point-rounded ability estimates indicates that when the two score information functions appear to differ, i.e. greater information for integer scores in the middle of the range for Math/Grade 11 and Math/Grade 5, differences in s.e.'s are not great. Differences in s.e.'s between 425 and 625 for Math/Grade 11 and between 450 and 600 for Math/Grade 5 are most frequently only one or two scale score points.

Nonnegligible differences in the precision of integer-rounded versus half-point-rounded number correct (# correct) ability estimates are limited to the lower part of the ability range. Integer-rounded number correct ability estimates, like their pattern counterparts, have substantially less error than half-point number correct ability estimates for this particular subrange of the Reading/Grade 11 test. At the floor the integer-rounded s.e. is smaller than the half-point-rounded s.e. by 124 points (212 versus 336; almost twice the approximately 64 point half-point and integer-rounded number correct standard deviations - Table 3) and is still seven points less at 425.

Half-point-rounded number correct estimates have marginally *less* error than integer-rounded estimates in the lower subranges for Science/Grade 11 and Math/Grade 5. The half-point s.e. of 112 at 300 is 17 points less than the integer-rounded s.e. of 129 for the Science test. The difference is reduced to three points by 375, however. A difference of 12 scale score points at the floor

of the Math/Grade 5 test (116 for half-point versus 128 for integer) has similarly been reduced to three points at 375.

Predicted Percentage of Maximum Score

Pattern Ability Estimates

The predicted pm's are provided at the selected scale score points for the four combinations of composite scores in Figures 1 through 4. Integer-rounded pm's tend to be slightly larger than half-point-rounded pm's, with most of the differences less than the 2.2 percentage point difference found at 500 to 525 for Science/Grade 11. Predicted pm's are actually slightly *smaller* (largest difference of .2) for integer-rounded ability estimates in the lower 300 to 375 subrange for Math/Grade 11 (e.g. 18.3 for integer-rounded c.r. scores at 325 versus 18.5 for half-point-rounded c.r. scores). The latter exception demonstrates that the effect of rounding to an integer does not necessitate a positive bias on test scores throughout the scale score range.

Number Correct Ability Estimates

Differences between predicted integer versus half-point-rounded c.r. scores are larger, though again differences are not invariably in favor of the integer scores. The largest bias is 7.8 percentage points found at 500 scale score points for Math/Grade 11. Above 600 to the ceiling of 700 for the Math/Grade 5 test integer-rounded predicted pm's are .2 to .3 *smaller* than half-point-rounded predicted pms's (e.g. 90.9 versus 91.1 at 675).

Comparisons of Ability Estimates within Examinees

Differences in scale score ability estimates produced for each examinee by integer versus half-point rounding were evaluated at each possible half point difference between the total raw c.r. scores (integer minus half-point). The range of *possible* differences in the total c.r. scores produced through integer-rounding and half-point rounding could vary between -.5 and +.5 times the number of c.r. items (including the null or zero difference). Not all possible differences were observed for each test. Differences for each type of ability estimate were evaluated.

Pattern Ability Estimates

Table 2 contains mean pattern scale score differences for various differences in total c.r scores. For the Reading/Grade 11 test, only two out of the three possible differences that could occur with a test containing a single c.r. item actually occurred: 0 and +5. The overall or sample mean difference at the bottom of the mean difference column was .05 (s.d. = 1.70). The mean difference between scale score estimates based on the two types of rounding (again integer minus half-point) for the 336 examinees who attained a +.5 difference in the total c.r. scores was 1.78 with the largest difference being 25 and the smallest difference being -.5 scale score points. These differences can be evaluated relative to the mean half-point pattern standard error for these 336 examinees: 20.32.

The other three tests have a larger number of differences between the total c.r. raw scores, as expected given the six to 11 c.r. items in these tests. The distributions of accumulated half point differences varies over the three tests, with the Science/Grade 11 test most asymmetric in having seven positive differences (every half point from .5 to 3.5) versus only two negative differences (-.5 and -1.0). The two Math tests are similar in having more approximately equal numbers of positive and negative differences, with the Grade 11 test having a wider range of differences (all possible half point differences for the six c.r. item test).

Distributions for all four tests exhibit a large number of zero differences in the total c.r. scores. The preponderance of differences are positive which is expected given the frequent rounding up, from a half point score to an integer, of an average score obtained when two readers differed by a point. The percentages of the four total samples that demonstrate negative total c.r. differences (i.e those students having at least one more average c.r. item score that is larger when rounded to the half point than when rounded to the integer) ranges from a low of 0% for Reading/Grade 11 to a maximum of 13% for Math/Grade 5 (11% for -.5 plus 2% for a -1.0 total c.r. difference).

Similar to the Reading/Grade 11 test, the overall mean differences for the other three forms were small, ranging between -.24 for Science/Grade 11 through .82 for Math/Grade 5. Mean scale score differences at each difference in the total c.r.

scores are not large but generally increase (decrease) with increases (decreases) in the difference in total c.r scores (e.g. from 2.20 to 6.17 as the difference in total c.r. score increases from .5 to 2.0 for Math/Grade 5).

Minimum or maximum within-examinee differences in scale score pattern estimates can be as large (in absolute value) as 35.5 at a total c.r. difference of 1.5 for Science/Grade 11. This value is more than one half of a pattern half-point (ph)or integer standard deviation (61.06 versus 61.11) and more than twice the mean ph standard error at that total c.r difference (16.10).

Number Correct Ability Estimates

Within-subject differences in number correct estimates in Table 3 are larger than the differences in pattern estimates but again, not large on average. Sample mean differences range between -.35 for Science/Grade 11 and 1.30 for Math/Grade 5. The largest maximum or minimum difference in number correct scale scores is -59 at a total c.r. score difference of -1.0 for Math/Grade 11.

DISCUSSION

The lack of a substantive improvement in measurement precision that could be attributed to half-point rounding, coupled with the documented instance of increased error induced by that type of rounding in a portion of the ability range of students taking one test (Reading/Grade 11), would seem to argue for rounding average c.r. ratings to the nearest integer. It can not

be assumed that half point c.r. scores will always meaningfully discriminate examinees on the ability being assessed. Rounding up to the nearest integer gives a preponderance of students the "benefit of the doubt" concerning the acceptability of their response. That is, those students obtaining two readings that differ by a point, consequently receiving a half point average score (e.g. 1.5 or 2.5), are awarded the greater integer score. Those students that require a third reading of their response and obtain an average with a remainder of 2/3's (e.g. three readers specifying a 0, 2 and 3 that averages 1.67) will obtain the closest integer to their unrounded score.

Finally, the use of integer-rounded as opposed to half-point-rounded c.r. scores has the important advantage of ensuring that final c.r. scores can be interpreted relative to specified levels of the item rubrics. A meaning of a half-point c.r. score, even if it served to discriminate examinees on the trait, would have to be "interpolated" between the rubric levels.

A decision to round average scores to the nearest integer would, however, result in a relatively small percentage of the examinee population (under 15%, given tests similar in the relative proportion of c.r. items to those studied) obtaining a scale score that was, on average, slightly reduced relative to what would be obtained if rounding to the half point occurred. The largest *average* mean scale score reduction for a *group* of students that would score lower with integer-rounded c.r. scores (those having negative differences in the total c.r. scores) was

-9.93 for pattern ability estimates and -13.43 for number correct estimates, although individual decreases as large as -30.0 and -59 scale score points were noted for these two types of estimates, respectively.

These latter two differences, while large relative to the sample scale score standard deviations, are not substantially larger than their standard errors. Only one student (taking the Math/Grade 5 test) across all four test samples that had a negative total c.r. difference also had a difference in pattern scale score estimates that exceeded one (half-point-rounded) s.e.. A maximum of eight students in one test sample (Math/Grade 11) had a difference in number correct scale scores that was in excess of one (half-point) number-correct s.e. with five in Math/Grade 5 and two in Science/Grade 11 also having differences larger than a standard error.

In terms of raw score points examinees doing worse under integer-rounded scoring lose 1/3 of a raw score point for every -.5 difference in total c.r. scores. This occurs when they have an average rating for a c.r. item with a 1/3 remainder that gets rounded down to the lower integer instead of to the closer half point. Consequently the single student who attained a -3.0 difference in total c.r. scores for Math/Grade 11 gives up the most in raw score points by rounding to integers (1/3 times six half points or 2 raw score points), though this student's difference in pattern scale scores is a relatively small 12 scale score points.

It should be noted that a policy of rounding to half score points instead of integers results in reductions, albeit smaller in magnitude, in raw score points for some examinees. These students are those that obtained an average score with a 2/3's remainder and lose the difference of approximately .17 between 2/3's and ½ that accompanies half-point rounding.

The number of students incurring a reduction in raw score due to either integer-rounding or half-point-rounding (or the magnitude of the reductions) can not be determined from Tables 2 and 3. A half-point difference between an average item score rounded to an integer could have been compensated for by a half point loss due to the other type of rounding (excluding those examinees who obtained the maximum or minimum possible difference in total c.r. scores). If the probability the rating process produced an average score with a remainder of 1/3 was the same as that of producing an average score with remainder of 2/3's there would be as many instances of losses due to half-point as integer-rounding. Hence the magnitude of the raw score reduction, summed over examinees, for half-point rounding would be twice that for integer-rounding (approximately .33 times the number of students impacted versus approximately .17 times a putative equal number of students) Unfortunately some of the tests studied here contradict that rating process assumption (e.g. Reading/Grade 11).

CONCLUSIONS

- There was little substantive difference in score information or the s.e.'s of ability estimates due to type of rounding, integer versus half-point, above the floors of three out of the four tests studied.
- In the fourth, Reading/Grade 11 test there was decidedly less error (more precision) in the integer-rounded ability estimates at the lower portion of the ability continuum (from a scale score of 300 to approximately 425). This was true for both pattern and number correct estimates.
- Integer-rounded estimates generally produce slightly larger predicted percent of maximum (test) scores, though not throughout the entire ability range of all of the four test studied. The expected larger positive differences or rounding bias for number correct estimates were observed.
- Within-subject differences between scale score estimates derived using integer versus half-point scores were generally small for both pattern and number correct ability estimates. Several differences between approximately 29 and 36 scale score points for pattern ability estimates and between 50 and 60 points for number correct ability estimates were observed, however. Differences this large were approximately one half and one standard deviation of the respective sample standard deviations.
- For those students scoring higher with half-point-rounded scores, a very small number had within-subject differences in

integer-rounded versus half-point rounded pattern or number correct ability estimates that were as large as one standard error in magnitude. None were as large as two s.e.'s.

References

Burket, G.R. (1991; 1995). *PARDUX*. Monterey, CA: CTB/McGraw-Hill.

Fitzpatrick, A.R. (1990). Status report on the results of preliminary analyses of dichotomous and multi-level items using the PARMATE (PARDUX) program. Unpublished manuscript.

Fitzpatrick, A.R., Link, V.B., Yen, W.M., Burket, G.R., Ito, K., & Sykes, R.C. (1996). Scaling performance assessments: A comparison of one-parameter and two-parameter partial credit models. *Journal of Educational Measurement, 33*, 291-314.

Lord, F.L. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Lawrence Erlbaum associates.

Masters, G.N. (1982). A Rasch model for partial credit scoring. *Psychometrika, 47*, 149-174.

Muraki, E. (1992). A generalized partial credit model: Applicationof an EM algorithm. *Applied Psychological Measurement, 16*, 159-176.

Thissen, D. (1986). *MULTILOG: Multiple categorical item analysis and test scoring, Version 5*. Mooresville, IN: Scientific Software.

Yen, W.M. (1981). Using simulation results to choose a latent trait model. *Applied Psychological Measurement, 5*, 245-262.

Table 1
Descriptive Statistics

Type of Rounding	# of Scored Items	# Points Total [cr]	% pts from c.r. items	Mean Total Score	s.d.	c.r. % of Max Score[1]	c.r. Mean score	c.r. s.d.	N	α	p-value[2] Mean	p-value[2] s.d.	Item - Test Correlation Mean	Item - Test Correlation s.d.
Reading/Grade 11														
Half-Point	36	38 [3]	8	21.60	7.22	57	0.66	0.93	1367	0.86	0.59	0.16	0.42	0.11
Integer				21.73	7.28	57	0.79	1.03		0.86	0.59	0.16	0.42	0.11
Science/Grade 11														
Half-Point	50	60 [18]	30	32.20	11.23	54	6.87	4.45	1293	0.90	0.57	0.16	0.41	0.12
Integer				32.99	11.45	55	7.67	4.73		0.90	0.57	0.16	0.41	0.12
Math/Grade 11														
Half-Point	46	61 [21]	34	30.67	12.74	50	6.19	5.67	1208	0.90	0.57	0.19	0.45	0.12
Integer				31.11	12.83	51	6.63	5.75		0.91	0.57	0.19	0.45	0.13
Math/Grade 5														
Half-Point	60	75 [26]	35	26.12	13.07	35	6.89	5.39	1541	0.91	0.38	0.21	0.40	0.14
Integer				26.36	13.13	35	7.13	5.47		0.91	0.38	0.21	0.40	0.14

[1] Mean divided by maximum score (percentage of maximum score)

[2] p-Value for constructed response items computed as average score divided by the maximum score

Table 2
Difference in Scale Scores by Differences in Total CR Score due to Type of Rounding
Pattern Scoring

Difference in Total CR Score (Integer - half)	Reading/Grade 11					Science/Grade 11					Math/Grade 11					Math/Grade 5				
		Differences (integer - half)			Mean s.e.		Differences (integer - half)			Mean s.e.		Differences (integer - half)			Mean s.e.		Differences (integer - half)			Mean s.e.
	N (%)	Min	Mean	Max	(PH)	N (%)	Min	Mean	Max	(PH)	N (%)	Min	Mean	Max	(PH)	N (%)	Min	Mean	Max	(PH)
-3											1 (0%)	-12.0	**-12.00**	-12.0	18.00					
-2.5											1 (0%)	-8.0	**-8.00**	-8.0	12.00					
-2											7 (1%)	-10.5	**-7.29**	-4.5	13.57					
-1.5											11 (1%)	-7.0	**-5.64**	-4.0	13.45	4 (0%)	-4.0	**-3.63**	-2.5	11.75
-1						7 (1%)	-12.5	**-9.93**	-8.5	15.14	37 (3%)	-30.0	**-5.28**	-2.5	15.22	24 (2%)	-24.0	**-4.31**	-1.0	13.50
-0.5						9 (1%)	-11.5	**-7.06**	-5.0	14.67	40 (3%)	-7.0	**-3.11**	0.0	13.50	172 (11%)	-29.5	**-2.58**	0.5	14.53
0	1031 (75%)	-5.5	**-0.51**	4.0	26.28	345 (27%)	-19.5	**-3.13**	23.5	23.82	436 (36%)	-20.5	**-1.28**	21.5	23.01	619 (40%)	-5.0	**-0.13**	6.5	21.05
0.5	336 (25%)	-0.5	**1.78**	25.0	20.32	332 (26%)	-6.5	**-1.81**	12.5	16.65	340 (28%)	-2.5	**0.44**	23.0	15.40	507 (33%)	0.0	**2.20**	23.5	15.97
1						270 (21%)	-4.0	**0.43**	18.5	15.90	192 (16%)	-2.0	**1.61**	16.0	14.57	171 (11%)	2.0	**3.39**	13.0	13.16
1.5						192 (15%)	-1.5	**3.05**	35.5	16.10	79 (7%)	0.0	**3.27**	29.5	14.06	41 (3%)	3.0	**4.68**	10.0	12.78
2						82 (6%)	0.5	**4.62**	23.0	15.23	44 (4%)	2.0	**4.02**	8.5	13.73	3 (0%)	6.0	**6.17**	6.5	11.67
2.5						41 (3%)	3.5	**7.13**	27.5	15.17	9 (1%)	2.0	**4.72**	10.0	14.22					
3						12 (1%)	4.0	**8.38**	13.0	14.75	11(1%)	3.5	**5.64**	7.0	12.64					
3.5						3 (0%)	10.5	**11.67**	13.5	14.67										
Sample Mean Difference			**0.05**					**-0.24**					**-0.01**					**0.82**		
s.d.			**1.70**					**4.60**					**3.36**					**2.78**		
Scale Score Means (s.d.'s)																				
Total N	1,367					1,293					1,208					1,541				
Pattern halfpoint	488.27 (64.19)					486.86 (61.06)					494.60 (60.47)					485.17 (60.72)				
Pattern integer	488.32 (64.11)					486.62 (61.11)					494.59 (60.12)					485.99 (61.09)				

PH = Pattern halfpoint

BEST COPY AVAILABLE

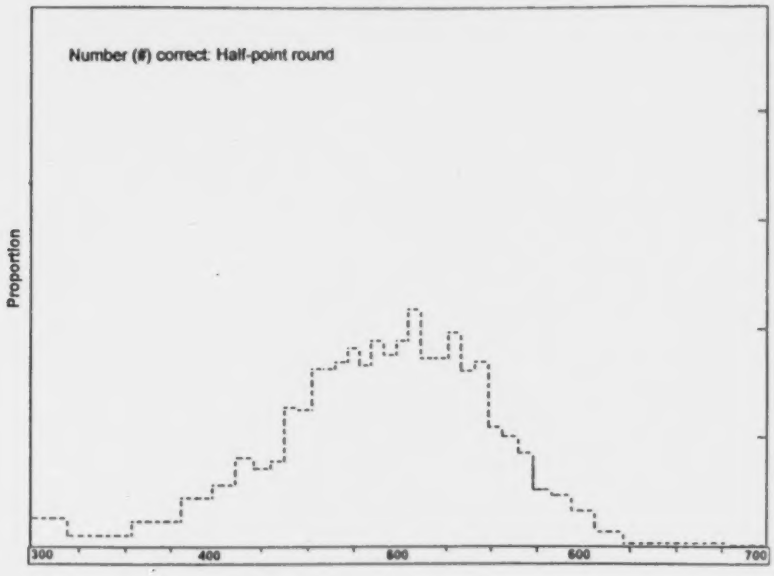
Table 3
Difference in Scale Scores by Differences in Total CR Score due to Type of Rounding
Number Correct Scoring

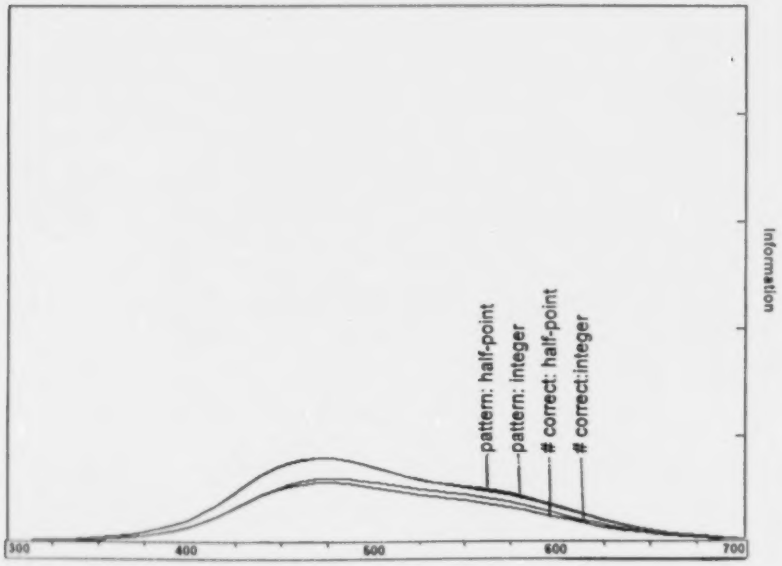
Difference in Total CR Score (Integer - half)	Reading/Grade 11					Science/Grade 11					Math/Grade 11					Math/Grade 5				
		Differences (integer - half)					Differences (integer - half)					Differences (integer - half)					Differences (integer - half)			
	N (%)	Min	Mean	Max	Mean s.e. (PH)	N (%)	Min	Mean	Max	Mean s.e. (PH)	N (%)	Min	Mean	Max	Mean s.e. (PH)	N (%)	Min	Mean	Max	Mean s.e. (PH)
-3											1 (0%)	-31	**-31.00**	-31	19.00					
-2.5											1 (0%)	-18	**-18.00**	-18	12.00					
-2											7 (1%)	-19	**-11.14**	-5	16.57					
-1.5											11 (1%)	-22	**-11.00**	-5	16.64	4 (0%)	-7	**-3.00**	4	13.50
-1						7 (1%)	-23	**-13.43**	-5	16.14	37 (3%)	-59	**-10.03**	4	18.43	24 (2%)	-31	**-6.66**	7	15.54
-0.5						9 (1%)	-15	**-7.44**	4	15.56	40 (3%)	-53	**-9.35**	0	17.03	172 (11%)	-26	**-3.35**	14	16.69
0	1031(75%)	-24	**0.11**	31	29.29	345 (27%)	-46	**-0.85**	23	26.86	436 (36%)	-26	**1.32**	17	30.03	619 (40%)	-31	**1.50**	33	25.48
0.5	336 (25%)	-11	**0.41**	31	23.56	332 (26%)	-33	**-2.62**	16	18.23	340 (28%)	-17	**0.17**	12	19.59	507 (33%)	-32	**2.18**	18	18.65
1						270 (21%)	-28	**-0.51**	40	17.24	192 (16%)	-22	**1.65**	34	17.40	171 (11%)	-10	**3.01**	16	15.22
1.5						192 (15%)	-16	**2.05**	50	17.37	79 (7%)	-9	**4.77**	56	17.48	41 (3%)	-13	**4.37**	15	14.88
2						82 (6%)	-16	**2.93**	33	16.30	44 (4%)	-4	**5.82**	19	16.70	3 (0%)	3	**7.00**	10	13.67
2.5						41 (3%)	-9	**5.71**	29	16.17	9 (1%)	-1	**10.89**	18	18.56					
3						12 (1%)	-4	**7.00**	20	15.50	11(1%)	6	**11.27**	17	15.36					
3.5						3 (0%)	9	**17.00**	29	15.33										
Sample Mean Difference			**0.19**					**-0.35**					**0.67**					**1.30**		
s.d.			**5.67**					**9.08**					**7.86**					**6.93**		
Scale Score Means (s.d.'s)																				
Total N	1,367					1,293					1,208					1,541				
Number correct half-point	488.11 (64.22)					486.87 (61.08)					493.40 (62.92)					485.46 (60.96)				
Number correct integer	488.29 (63.71)					486.52 (61.38)					494.07 (61.33)					486.76 (60.40)				

NH = Number half-point

BEST COPY AVAILABLE

Figure 1
Reading/Grade 11



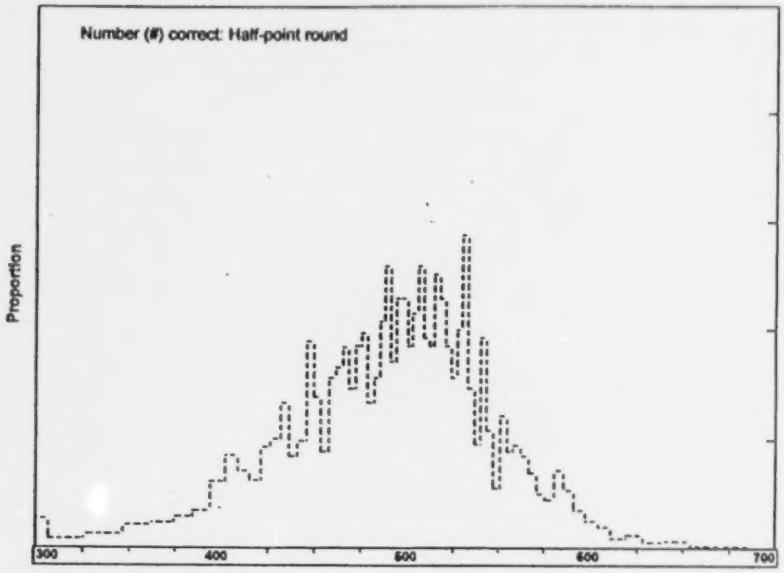


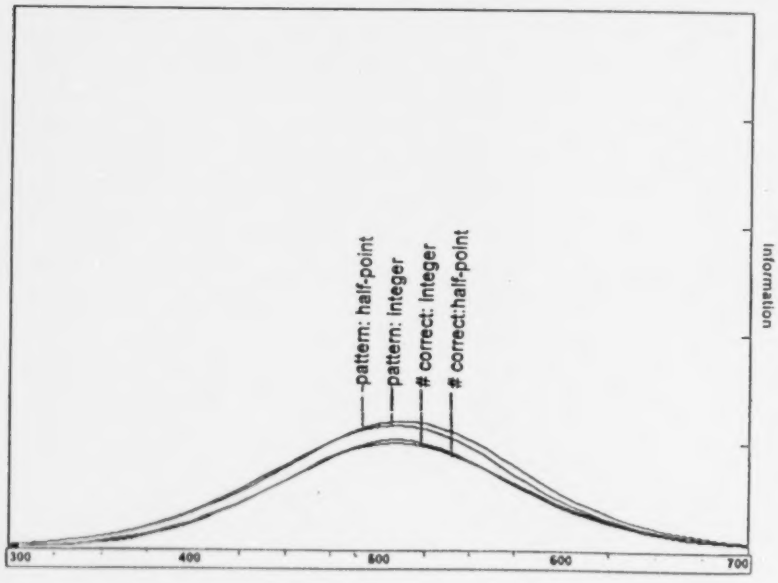
Scale Score	Integer # Correct Predicted PM	s.e.	Integer Pattern Predicted PM	s.e.	Half-Point # Correct Predicted PM	s.e.	Half-Point Pattern Predicted PM	s.e.
300	19.4	212	20.1	132	18.1	336	20.0	201
325	20.3	137	20.8	95	18.9	216	20.7	142
350	21.7	88	21.9	67	20.2	136	21.8	99
375	23.9	57	23.9	46	22.4	87	23.8	69
400	27.7	36	27.7	31	25.9	53	27.5	45
425	33.7	25	34.3	21	31.5	32	34.0	27
450	42.1	20	44.1	17	39.4	22	43.7	18
475	51.9	19	55.1	17	48.7	19	54.5	16
500	61.4	20	65.2	18	58.0	19	64.5	16
525	70.1	21	73.7	19	66.7	20	73.0	17
550	77.8	22	81.1	21	74.8	22	80.4	19
575	84.4	25	87.3	23	82.1	25	86.8	20
600	89.5	29	92.0	26	87.9	29	91.7	24
625	93.0	36	95.2	33	91.9	37	95.0	30
650	95.3	46	97.1	42	94.5	48	96.9	40
675	96.8	60	98.2	55	96.1	62	98.0	53
700	97.7	76	98.8	70	97.1	79	98.7	69

PM = Percentage of Max Score

BEST COPY AVAILABLE

Figure 2
Science/Grade 11



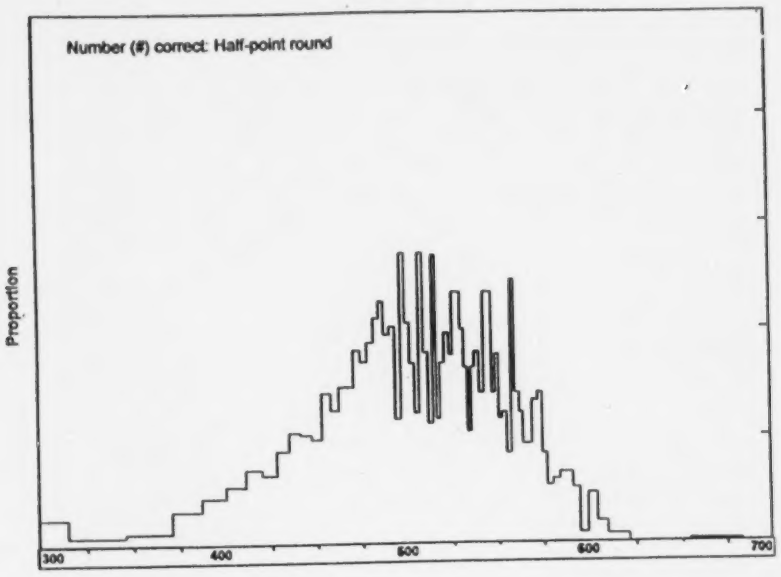


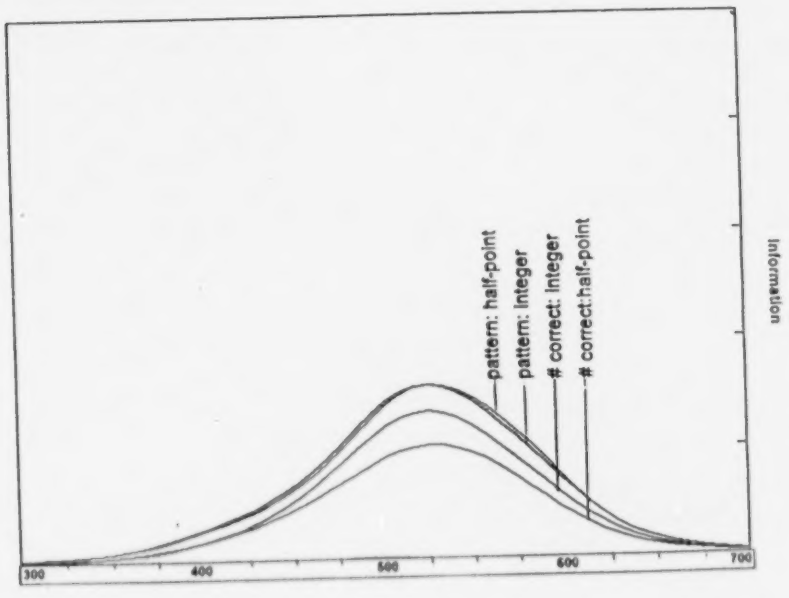
	Integer				Half-Point			
	# Correct		Pattern		# Correct		Pattern	
Scale	Predicted		Predicted		Predicted		Predicted	
Score	PM	s.e.	PM	s.e.	PM	s.e.	PM	s.e.
300	17.5	129	18.9	77	13.7	112	18.3	73
325	18.6	85	19.8	56	14.7	76	19.3	55
350	20.4	57	21.4	41	16.3	52	20.7	41
375	23.0	40	23.8	31	18.7	37	23.0	32
400	27.0	29	27.5	24	22.2	28	26.5	25
425	32.6	22	32.8	19	27.3	22	31.5	20
450	40.0	18	40.1	16	34.2	18	38.4	17
475	49.0	16	49.3	15	43.1	16	47.2	15
500	59.2	15	59.8	14	53.5	15	57.6	14
525	69.2	15	70.4	14	64.2	15	68.2	14
550	77.9	16	79.7	15	74.0	16	77.7	15
575	84.7	19	86.7	18	81.5	19	85.1	17
600	89.6	23	91.5	22	87.3	23	90.3	21
625	92.9	29	94.5	27	91.3	28	93.6	26
650	95.2	36	96.5	34	94.0	34	95.8	32
675	96.8	44	97.7	42	95.8	42	97.2	39
700	97.8	55	98.5	53	97.1	51	98.1	48

PM = Percentage of Max Score

BEST COPY AVAILABLE

Figure 3
Math/Grade 11



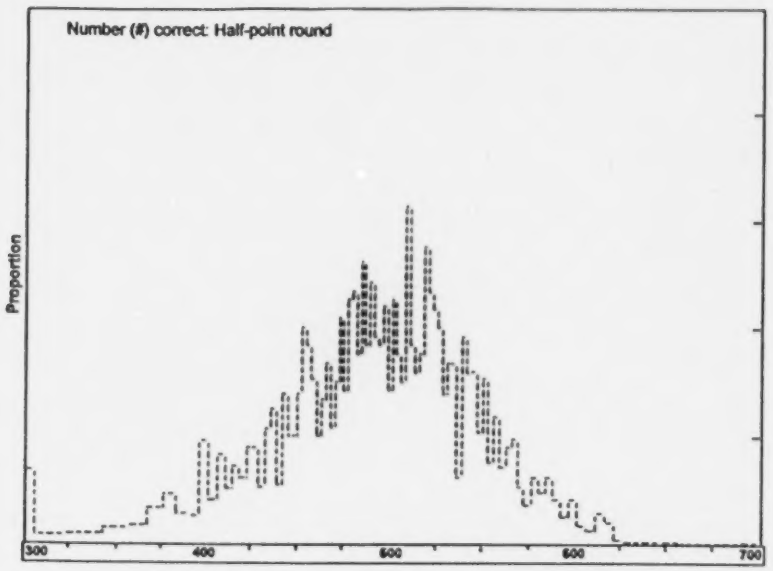


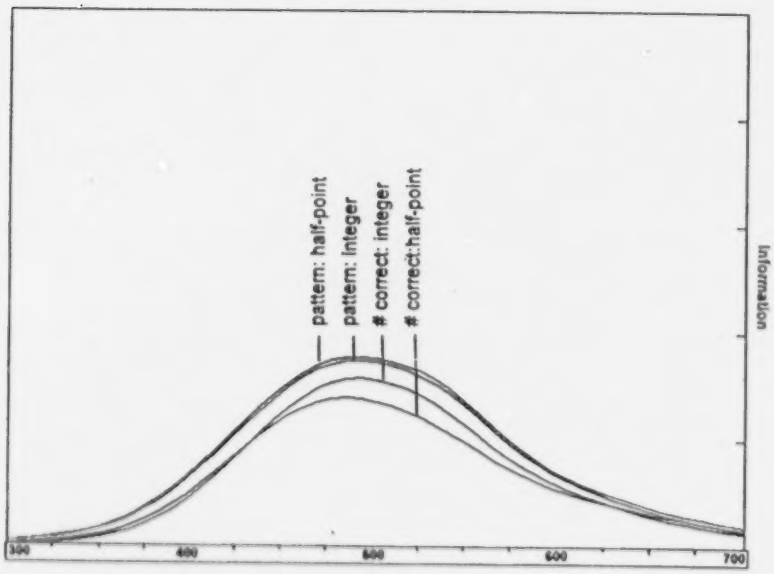
Scale Score	Integer				Half-Point			
	# Correct		Pattern		# Correct		Pattern	
	Predicted PM	s.e	Predicted PM	s.e	Predicted PM	s.e	Predicted PM	s.e
300	15.2	206	17.8	108	11.6	200	18.0	113
325	15.9	129	18.3	77	12.2	128	18.5	82
350	17.0	81	19.3	53	13.1	82	19.4	57
375	18.8	52	20.9	37	14.5	54	21.0	40
400	21.5	36	23.6	28	16.8	37	23.6	29
425	25.6	26	27.7	22	20.1	28	27.6	23
450	31.3	20	33.6	18	24.9	22	33.2	18
475	39.2	16	41.9	14	31.9	18	41.3	15
500	49.6	13	53.0	12	41.8	15	52.1	12
525	62.0	13	65.6	12	54.9	14	64.6	12
550	74.4	14	77.4	12	69.2	15	76.6	12
575	84.6	16	86.7	14	81.6	17	86.3	14
600	91.5	20	93.0	18	90.0	23	92.9	17
625	95.3	29	96.5	26	94.5	32	96.4	25
650	97.3	42	98.2	37	96.8	46	98.1	37
675	98.3	59	99.0	53	97.9	64	98.9	53
700	98.9	82	99.4	73	98.5	85	99.3	72

PM = Percentage of Max Score

BEST COPY AVAILABLE

Figure 4
Math/Grade 5





Scale Score	Integer				Half-Point			
	# Correct		Pattern		# Correct		Pattern	
	Predicted		Predicted		Predicted		Predicted	
	PM	s.e	PM	s.e	PM	s.e	PM	s.e
300	7.9	128	7.7	77	6.1	116	7.6	79
325	8.6	91	8.2	60	6.7	84	8.1	61
350	9.6	65	9.0	46	7.6	60	8.9	47
375	11.1	47	10.3	36	9.0	44	10.1	36
400	13.3	34	12.3	28	11.0	33	12.1	28
425	16.6	26	15.3	22	14.0	26	15.2	22
450	21.2	20	19.9	18	18.4	21	19.7	18
475	27.4	17	26.2	15	24.6	17	26.1	15
500	35.6	14	34.6	13	33.0	15	34.9	13
525	45.6	13	45.7	12	43.6	13	45.9	11
550	56.7	12	58.0	11	55.6	13	58.4	11
575	67.3	13	69.7	12	66.9	14	70.1	12
600	76.0	15	79.0	14	76.0	16	79.3	14
625	82.6	18	85.7	16	82.8	19	85.9	17
650	87.4	21	90.5	20	87.7	23	90.4	20
675	90.9	26	93.7	24	91.1	28	93.6	24
700	93.2	34	95.8	30	93.5	34	95.6	30

PM = Percentage of Max Score

BEST COPY AVAILABLE